Western Alaska Salmon Stock Identification Program **Technical Document**:[1] 10

Title: Optimal Rate of Correct Assignment with backward elimination locus selection **Version:** 1.0

Authors: J. Jasper and W. Templin
Date: December 14, 2010

Introduction

As part of the locus selection process proposed for chum salmon in WASSIP, we propose using f_{ORCA} (Rosenberg et al. 2003; Rosenberg 2005) with backward elimination as one of the marker selection methods for choosing SNPs for the chum salmon baseline (Tech Doc 8). Results from this analysis are proposed to provide 30% of the locus-selection weight, the most of any analysis. The information measure, f_{ORCA}, returns the Optimal Rate of Correct Assignment (ORCA) for a particular locus set with respect to a specific baseline. At each iteration of the routine, a randomly drawn individual is assigned to a population for which its genotypic probability is a maximum. We propose adapting f_{ORCA} to allow us to determine the best set of loci to provide separation among reporting groups taking advantage of potential synergy among loci. To do this we propose implementing a backward elimination algorithm similar to that described in BELS (Bromaghin 2008). However, we opted not to use the program BELS because it is too time-consuming. Even though the Gene Conservation Laboratory does proportional allocation (as does BELS) rather than individual assignment (as does f_{ORCA}), we feel that f_{ORCA} with backward elimination has merit under a Bayesian mixed stock analysis routine because it attempts to select a suite of markers that optimizes the genotypic probabilities of potential mixture individuals, and BAYES (Pella and Masuda 2001) uses these probabilities to stochastically assign the mixture individuals each iteration.

Current f_{ORCA} Algorithm

While a closed form solution of f_{ORCA} is available (Rosenberg et al. 2003), it becomes impractical for large locus sets. Therefore, Rosenberg (2005) provided an iterative algorithm for estimating f_{ORCA}. This algorithm can be explained as follows.

1. Uniformly draw a population at random from the baseline.
2. Randomly generate a multi-locus genotype based on the allele frequencies of the population chosen in the first step.

[1] This document serves as a record of communication between the Alaska Department of Fish and Game Commercial Fisheries Division and the Western Alaska Salmon Stock Identification Program Technical Committee. As such, these documents serve diverse ad hoc information purposes and may contain basic, uninterpreted data. The contents of this document have not been subjected to review and should not be cited or distributed without the permission of the authors or the Commercial Fisheries Division.

3. Assign that genotype to the population for which its genotypic probability is a maximum.
4. Repeat Steps 1-3 10,000 times.
5. After repeating this process multiple times, f_{ORCA} is calculated as the proportion of times that the assignment in Step 3 is the same population drawn in Step 1.

While f_{ORCA} is typically used to evaluate how well a marker set can assign individuals back to the correct population, it could also be adapted for evaluating how well a marker set can be used to assign individuals back to the correct region. With this application the algorithm would be as follows.

1. Uniformly draw a population at random from the baseline.
2. Determine the region to which the population belongs.
3. Randomly generate a multi-locus genotype based on the allele frequencies of the population chosen in the first step.
4. Assign that genotype to the population for which its genotypic probability is a maximum.
5. Determine the region to which the assignment population belongs.
6. Repeat Steps 1-5 10,000 times.
7. After repeating this process multiple times, f_{ORCA} is calculated as the proportion of times that the assignment in Step 5 is the same region drawn in Step 2.

Backward Elimination Locus Selection Algorithm

Rosenberg's f_{ORCA} algorithm provides a means of evaluating the performance of a locus set, but it does not provide us with an algorithm for selecting sets of markers to evaluate. Rosenberg (2005) does provide four such algorithms and discusses the advantages and limitations of each: 1) Exhaustive evaluation, 2) Univariate accumulation, 3) Greedy accumulation, and 4) Maxmin accumulation.

One locus selection algorithm that Rosenberg failed to discuss is the method used in the Backward Elimination Locus Selection (BELS) algorithm laid-out by Bromaghin (2008). This algorithm has the advantages of being both simple to implement and it exploits synergies among loci. However, Bromaghin (2008) does not use f_{ORCA} to evaluate marker sets; rather he uses actual maximum likelihood mixed stock analysis and bootstrap simulations to evaluate performance in the software BELS. While we agree that this is a relevant measure, unlike f_{ORCA}, it suffers from being prohibitively slow and may be biased in some circumstances (Anderson 2008).

We suggest that marker selection applications with large numbers of populations and loci should employ the BELS algorithm for selecting marker panels to evaluate, but use the f_{ORCA} function to do the evaluation. For the purposes of WASSIP, we will use the correct assignment to region algorithm described above.

This would be accomplished by the following:

1. Start with entire set of L potential markers.
2. Create L sub-sets of L-1 markers by removing each marker, in turn, from full the set.
3. Evaluate f_{ORCA} on all L sub-sets using correct assignment to region.
4. Identify sub-set with maximum f_{ORCA}.
5. Record which locus was removed.
6. Return to Step 1 using the sub-set identified in Step 4 as the new full set of L-1 loci.

This process is continued until no markers remain. The loci can be ranked according to the order in which they were removed or scored according to their f_{ORCA} value.

This algorithm has been implemented in R for use with the chum salmon SNP selection process described in Technical Document 8, "Chum salmon SNP selection process outline."

The limitations of f_{ORCA} are: 1) it (likely) suffers from providing an optimistic rate of correct assignment, and; 2) spurious differences in allele frequencies can lead to falsely identifying some loci as influential. An extension of f_{ORCA} that may alleviate its limitations would be to implement a "leave-one-out" approach by which we randomly draw an individual from the ascertainment baseline, recalculate the allele frequencies without that individual, then assign the individual based on the recalculated allele frequencies. While more difficult to implement, this version may be a more viable solution. We are currently working on programming this extension.

Citations

Anderson E.C., R.S. Waples, S.T. Kalinowski. 2008. An improved method for estimating the accuracy of genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences* 65:1475-1486.

Bromaghin, JF. 2008. BELS: backward elimination locus selection for studies of mixture composition or individual assignment. *Molecular Ecology Resources* 8: 568-571

Rosenberg, NA, LM Li, R Ward, & JK Pritchard. 2003. Informativeness of Genetic Markers for Inference of Ancestry. *American Journal of Human Genetics* 73 (1421):1402-1422

Rosenberg, NA. 2005. Algorithms for Selecting Informative Marker Panels for Population Assignment. *Journal of Computational Biology* 12 (9):1183–1201

Technical Committee review and comments

General comments: In general the approach seems reasonable, but we have some specific comments as detailed below.

Minor comments:

Line 13: "At each iteration of the routine, a randomly drawn individual is assigned to a population for which its genotypic probability is a maximum." How is this individual chosen? What is the pool of candidate individuals?

Line 29: "Uniformly draw a population at random from the baseline." What exactly does this mean? Each population has equal weight, and then the draw is random?

Line 63: "While we agree that this is a relevant measure, unlike *fORCA*, it suffers from being prohibitively slow and may be biased in some circumstances (Anderson 2008)." After "unlike *fORCA",* two attributes are listed but only one (being slow) is unlike *fORCA.* The bias described by Anderson et al. (2008) is equally applicable to *fORCA.* See below for more on this point.